A greedy dimension reduction method for classification problems

Damiano Lombardi, Fabien Raphel

A GREEDY DIMENSION REDUCTION METHOD FOR CLASSIFICATION PROBLEMS

DAMIANO LOMBARDI* AND FABIEN RAPHEL,†

Abstract. In numerous classification problems, the number of available samples to be used in the classifier training phase is small, and each sample is a vector whose dimension is large. This regime, called *high-dimensional/low sample size* is particularly challenging when classification tasks have to be performed. To overcome this shortcoming, several dimension reduction methods were proposed. This work investigates a greedy optimisation method that builds a low dimensional classifier input. Some numerical examples are proposed to illustrate the performances of the method and compare it to other dimension reduction strategies.

1. Introduction. This work investigates a method of dimension reduction applied to classification problems. These arise in many areas of applied sciences in which data are queried to provide predictions in a form of *yes/no* answers or more elaborated classification outcomes. Often, prior of classification, data are pre-processed in order to train in a more effective way a classifier. Part of the pre-processing phase takes the form of a linear or non-linear dimension reduction. Hereafter we propose a systematic way of performing this task.

Let G be an ensemble of signals, provided from experimental measurements, numerical simulations (or both). Let $n_s \in \mathbb{N}^*$ be the number of samples that will be used to train the classifier: for each $G^{(i)},\ i = 1, \dots, n_s$ a set of $n_g \in \mathbb{N}^*$ quantities are extracted from the signal. These can be either informed linear or non-linear forms identified by experimental insight or more agnostic features, such as point values of the signal, local average, Fourier or Wavelets coefficients. We refer to the set of these quantities for all the available signals as the dictionary entries $G^{(i)}_j \in \mathbb{R},\ i = 1, \dots, n_s,\ j = 1, \dots, n_g$. The present work deals with classification problems, namely, given an observable signal coming from a physical system, we want to determine to which class in a set of possible classes the system belongs to. In the present work, for sake of simplicity, the method is derived in the case of binary classification: its extension to multiple classes is straightforward. The methodology presented is general, and it was motivated by classification problems arising in biomedical engineering, in which the problems at hand can be sometimes in a different regime with respect to the ones classically addressed in Machine Learning. Indeed, as in other fields of science and engineering, the size n_g of quantities that can be extracted from the signal can be extremely large. Moreover, the number of available samples n_s, due to experimental constraints and to the complexity of the systems at hand, can be small if compared to n_g. This regime, called *high dimensional/low sample size* in the learning community, is particularly critic when performing classification and regression tasks. The mathematical reason is that we wish to identify a function whose domain dimension n_g is large, and hence we are exposed to the phenomenon of the *curse of dimensionality*, introduced for the first time by Bellman in [2] and related to learning theory in [39]. In [6, 15, 31] a theoretical analysis is proposed that describes the ability of approximating a high-dimensional ridge function by point queries and how the curse of dimensionality can be eventually circumvented. From a probabilistic viewpoint, for a given sample size, when the dictionary size becomes too large, the classification error increases: this is referred to as the Hughes

*Inria Paris, France (damiano.lombardi@inria.fr,fabien.raphel@inria.fr).
†NOTOCORD® part of Instem, Le Pecq, France (fabien.raphel@instem.com).

phenomenon [21, 40]. This regime is appearing in various areas of science and it is nowadays widely studied [12, 19, 29, 32].

To overcome this difficulty, a common strategy consists in reducing the dimension of the input space (consider [14] for an overview), that was considered in machine learning applications in [17, 20, 24, 27]. In most of the references, a dimension reduction strategy is applied and the results in terms of classification are then analysed. In the present work, we proposed to project the dictionary entries into a low dimensional linear subspace (obtained by a sparse linear combination of the entries), which is computed in order to optimize the classification success rate. From a dimension reduction point of view, the method proposed can be considered as a goal-oriented dimension reduction.

1.1. Notations and assumptions. Let $X_{n_g} \in \mathbb{R}^{n_g}$ be a random vector of the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We assume that the probability density function (pdf) of X_{n_g} is a mixture of the form:

$$\rho(g) = \rho_0(g)\pi_0 + \rho_1(g)\pi_1, \tag{1.1}$$

where $\rho_i(g) = \rho_i(g|y_* = i)$, the conditional probability density of g given that its label is $y_* = i$. The scalars π_i are the weights of the mixture and they can be seen as the *a priori* probability mass of being in the class i. It holds $\pi_0 + \pi_1 = 1$.

A classifier is defined in Definition 1.1 with $n = n_g$.

DEFINITION 1.1. *Let $g \in \mathbb{R}^n$ be an observation coupled with a label y. A binary classifier is a function $\mathcal{C}_n$ such that the following holds:*

$$\mathcal{C}_n : \mathbb{R}^n \to \{0, 1\}, \tag{1.2}$$

$$g \mapsto y. \tag{1.3}$$

Some geometrical notations are introduced. Let $k \leq n_g$. The Grassmann manifold Gr_{k,n_g} is the set of $k-$dimensional linear subspace of $\mathbb{R}^{n_g}$. The method proposed in the present work can be seen as an optimisation on the compact Stiefel manifold, denoted by $\mathcal{M}_{k,n_g}$, whose definition is recalled in Definition 1.2. An element of the Stiefel manifold will be denoted by M.

DEFINITION 1.2. *A Stiefel manifold $\mathcal{M}_{k,n_g}$ is a set of all the k-frames in $\mathbb{R}^{n_g}$:*

$$\mathcal{M}_{k,n_g} \triangleq \Big\{ Y = (Y_1, \ldots, Y_k), Y_i \in \mathbb{R}^{n_g} | Y_i^T Y_j = \delta_{ij}, \forall 1 \leq i, j \leq k \Big\}, \tag{1.4}$$

so that the elements of the compact Stiefel manifold are the $k \times n_g$ matrices with orthonormal columns. The Stiefel manifold $\mathcal{M}_{n_g,n_g} = O(n_g)$ is the orthogonal group. An element $R \in O(n_g)$ satisfies $R^T R = RR^T = I_{n_g}$. It can be seen, roughly speaking, as the concatenation of an element of the Stiefel manifold, and its orthogonal complement: $R = [M, M^\perp]$. Let us consider the endomorphism induced by R, and how the probability density ρ is transformed accordingly. A change of coordinates is applied to the expression in Equation 1.1, leading to:

$$\rho(\xi) = \rho_0(R\xi)\pi_0 + \rho_1(R\xi)\pi_1, \tag{1.5}$$

that holds since $\det(R) = 1$.

In the context of high dimensionality, low sample size, we assume that $k \ll n_g$ (e.g 1, 2 or 3).

2. Method. The method is detailed. An element M of a Stiefel manifold is used to reduce the input dimension: $x \in \mathbb{R}^k$ (the dimension k is, also, an outcome of the proposed method).

Let $\mathcal{C}_k$ be a classifier in the projected space of dimension k (see Section 1.1 and Definition 1.1 with $n = k$). It is defined as:

DEFINITION 2.1. *The classifier $\mathcal{C}_k$ in the subspace of dimension $k \ll n_g$ is defined as follow:*

$$\begin{aligned} \mathcal{C}_k : \mathbb{R}^k &\longrightarrow \{0, 1\} \\ x = M^T g &\mapsto y \end{aligned}, \tag{2.1}$$

where $g \in \mathbb{R}^{n_g}$ is an observation, $M \in \mathcal{M}_{k,n_g}$ and y is the label in the projected space.

The objective is to find $M \in \mathcal{M}_{k,n_g}$ which maximizes the success rate of the classifier $\mathcal{C}_k$. This has to be made more precise. In particular, an objective function is introduced, related to the classification success rate. This function could, in general, depend upon the classifier (defined by the function $\mathcal{C}$); in what follows we will propose an objective function that is intrinsically related to the ability of distinguishing between two classes, and that can be applied to all types of classifiers.

2.1. Classification score in the reduced space. The classification score is investigated and its relation to the score in the dictionary space is derived. Roughly speaking, reducing the dimension reduces also the amount of information the input carries about the classification output. This loss has to be quantified and minimised.

First, a consideration on the projected density on a Stiefel manifold element is presented, which will be used to derive the relationship between the classification score and the total variation in the reduced input space.

By the properties of orthogonality of the Stiefel manifold and its orthogonal complement, the pdf p in the projected space of dimension $k < n_g$ corresponds to the marginals of ρ (see Equation 2.2). Indeed, let $M \in \mathcal{M}_{k,n_g}$ and $R = [M, M^\perp]$. Let an input $x = M^T g$; we denote by $\xi \in \mathbb{R}^{n_g}$ the vector $\xi = R^T g$. It follows that $x = [\xi_1; \dots; \xi_k]$. Since R is an element of the orthogonal group, it holds:

$$p(x) = \int_{\mathbb{R}^{n_g - k}} \rho(\xi) d\xi_{k+1} \dots d\xi_{n_g}, \tag{2.2}$$

and hence:

$$p(x) = p_0(x)\pi_0 + p_1(x)\pi_1. \tag{2.3}$$

An important consequence is that p is a mixture of the same form as ρ, and, moreover, $p_i(x)$ is the conditional probability:

$$p_i(x) = p(x|y_* = i), \tag{2.4}$$

for $i = 0$ or 1.

The input space is subdivided into three distinct regions, in relation to what the classifier $\mathcal{C}_k$ (see 2.1) would provide, based on a probability argument. We denote by $S_0 \subseteq \mathbb{R}^k$, $S_1 \subseteq \mathbb{R}^k$ and $S_2 \subseteq \mathbb{R}^k$:

DEFINITION 2.2.

$$\begin{cases} S_0 \triangleq \{x = M^T g \in \mathbb{R}^k | \pi_0 p_0(x) > \pi_1 p_1(x)\} \\ S_1 \triangleq \{x = M^T g \in \mathbb{R}^k | \pi_1 p_1(x) < \pi_0 p_0(x)\} \\ S_2 \triangleq \{x = M^T g \in \mathbb{R}^k | \pi_0 p_0(x) = \pi_1 p_1(x)\} \end{cases} . \tag{2.5}$$

It follows that:

- $S_i \cap S_j = \emptyset, \forall i \neq j$.
- $\cup_{i=0}^{2} S_i = S \subseteq \mathbb{R}^k$.

Let (g, y_*) be a couple such that $g \in \mathbb{R}^{n_g}$ is an observation and $y_* \in \{0, 1\}$ the corresponding label (the true label). Let A_S be the ensemble of the success events, *i.e.* when the classifier $\mathcal{C}_k$ provides as result $y = y_*$. The set of success events can be defined as:

DEFINITION 2.3.

$$\begin{cases} A_{S_0} \triangleq \{y_* = 0 \wedge \pi_0 p_0 > \pi_1 p_1\} \\ A_{S_1} \triangleq \{y_* = 1 \wedge \pi_1 p_1 > \pi_0 p_0\} \\ A_{S_2} \triangleq \{y_* = 0, 1 \wedge \pi_0 p_0 = \pi_1 p_1\} \end{cases} , \tag{2.6}$$

And,

$$A_S \triangleq \cup_{i=0}^{2} A_{S_i}. \tag{2.7}$$

From the success events defined in Definition 2.3, the measure of the success events $\mu(A_S)$ can be obtained by quantifying the measure of the set:

$$\mu(A_S) = \int_{S_0} \pi_0 p_0(x) dx + \int_{S_1} \pi_1 p_1(x) dx + \frac{1}{2} \int_{S_2} p(x) dx, \tag{2.8}$$

The $\frac{1}{2}$ factor is justified by the fact that we expect to have half of the realizations to be well classified on S_2. This score is analogous to the excess risk measure proposed in [4].

2.1.1. Relation to the total variation. In order to quantify the success rate of the classification, distances or divergences between densities are commonly used. We denote by δ_{TV} the total variation [1, 34] (see Definition 2.4[1]). The total variation is a f-divergence [7] which is also a metric over the probability densities.

DEFINITION 2.4. *Let P and Q be two probability distributions on $(\Omega, \mathcal{A})$ (with Ω the sample space and $\mathcal{A}$ a σ-algebra) and p and q the corresponding pdf. Then, the total variation is:*

$$\delta_{TV}(P, Q) = \frac{1}{2} \int_{\Omega} |p(x) - q(x)| dx. \tag{2.9}$$

The pertinence of the total variation in relation to classification can be hinted by the following consideration. When the total variation is 0, the probability distributions corresponding to the two classes coincide almost everywhere. It means that for any observation (up to a zero measure set), we could attribute either 0 or 1 and

[1] This definition is a variant of the original definition of the total variation [1, 16].

no discrimination between the two classes would be possible based on a probability argument.

In the following of this paper, we make the hypothesis that the total variation between ρ_0 and ρ_1 is strictly positive, *i.e.* $\delta_{TV}(\rho_0, \rho_1) > 0$. In the case of binary classification, we also assume that $\min(\pi_0, \pi_1) > 0$.

We show hereafter that the measure of success Eq.(2.8) is related to the total variation between the densities p_0, p_1:

PROPOSITION 2.5. *Let $p(x)$ be defined as in Eq.(2.3)-(2.4), and the quantity $\mu(A_S)$ be defined as in Eq.(2.8). It holds:*

$$\frac{1}{2} + \min(\pi_0, \pi_1)\delta_{TV}(p_0, p_1) \leq \mu(A_S) \leq \frac{1}{2} + \max(\pi_0, \pi_1)\delta_{TV}(p_0, p_1), \tag{2.10}$$

Proof. A first relationships between the score and the densities is derived from the normalisation condition. As $p(x)$ is a density, it holds:

$$\int_S p(x)\ dx = 1 \ \Rightarrow \int_{S_0} \pi_0 p_0 + \pi_1 p_1\ dx + \int_{S_1} \pi_0 p_0 + \pi_1 p_1\ dx + \int_{S_2} \pi_0 p_0 + \pi_1 p_1\ dx = 1,$$

by the properties of the measures and the definition of $p(x)$. The terms of the definition of $\mu(A_S)$ are isolated, providing:

$$\mu(A_S) + \frac{1}{2}\int_{S_2} p(x)dx + \int_{S_0} \pi_1 p_1(x)dx + \int_{S_1} \pi_0 p_0(x)dx = 1. \tag{2.11}$$

Second, by adding and subtracting the same terms to the definition of the score, aiming at highlighting its relationship with the total variation we have:

$$\mu(A_S) = \int_{S_0} (\pi_0 p_0 - \pi_1 p_1)dx + \int_{S_1} (\pi_1 p_1 - \pi_0 p_0)dx + \int_{S_0} \pi_1 p_1 dx + \int_{S_1} \pi_0 p_0 dx + \frac{1}{2}\int_{S_2} p dx.$$

After some algebra, and by making use of the result in Eq.(2.11), we get:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big(\int_{S_0} (\pi_0 p_0 - \pi_1 p_1)dx + \int_{S_1} (\pi_1 p_1 - \pi_0 p_0)dx\Big). \tag{2.12}$$

It holds that on S_0 we have $\pi_0 p_0 - \pi_1 p_1 > 0$ and the converse holds on S_1, almost everywhere. Moreover, on S_2 it holds that $\pi_0 p_0 - \pi_1 p_1 = 0$. Henceforth:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big(\int_S |\pi_0 p_0 - \pi_1 p_1| dx\Big). \tag{2.13}$$

As $\pi_0 + \pi_1 = 1$, it holds:

$$\frac{1}{2} + \frac{\min(\pi_0, \pi_1)}{2}\Big(\int_S |p_0 - p_1| dx\Big) \leq \mu(A_S) \leq \frac{1}{2} + \frac{\max(\pi_0, \pi_1)}{2}\Big(\int_S |p_0 - p_1| dx\Big). \tag{2.14}$$

Which concludes the proof. ∎

Remark 1:. From Equation 2.13 obtained in the proof, we can directly see that $\frac{1}{2} \leq \mu(A_S) \leq 1$. The lower bound is attained when $S_2 = S$ ($\pi_0 p_0 = \pi_1 p_1$ almost everywhere on S). In that case, the scaled densities are equal a.e. and the probability of being in class 0 or 1 is $\frac{1}{2}$, which means that on average, half of the observations are well classified.

The result of the Proposition presented above states that the success rate of the classifier using the reduced input x can be directly related to the total variation of the projected densities. Aiming at quantifying the loss with respect to the classifier that exploits at best all the dictionary entries, we prove the following result:

PROPOSITION 2.6. *Let ρ_0, ρ_1 be the densities defined in Eq.(1.1). Then, it holds:*

$$\mu(A_S) \leq \frac{1}{2} + \max(\pi_0, \pi_1)\delta_{TV}(\rho_0, \rho_1), \tag{2.15}$$

Proof. Let M be an element of the Stiefel manifold $\mathcal{M}_{k,n_g}$, and $M^\perp$ its orthogonal complement of M. The score $\mu(A_S)$, by exploiting the change of coordinates and the properties of the elements of the orthogonal group, can be rewritten as follow:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\left(\int_{M_k}\left|\int_{M_k^\perp}(\pi_0\rho_0 - \pi_1\rho_1)d\xi_{k+1}\dots d\xi_{n_g}\right| d\xi_1\dots d\xi_k\right). \tag{2.16}$$

By triangular inequality, we can write:

$$\mu(A_S) \leq \frac{1}{2} + \frac{1}{2}\left(\int_{M_k}\int_{M_k^\perp}|\pi_0\rho_0 - \pi_1\rho_1|\, d\xi_{k+1}\dots d\xi_{n_g} d\xi_1\dots d\xi_k\right). \tag{2.17}$$

Finally, by exploiting the fact that $\pi_0 + \pi_1 = 1$:

$$\mu(A_S) \leq \frac{1}{2} + \max(\pi_0, \pi_1)\delta_{TV}(\rho_0, \rho_1). \tag{2.18}$$

∎

The result of the above Proposition shows that the total variation in the dictionary space of dimension n_g is a natural upper bound of the score. By projecting on a Stiefel manifold we cannot improve with respect to the best classifier that uses all the information.

An illustration of the relationships between the score and the total variation is proposed in Figure 1.

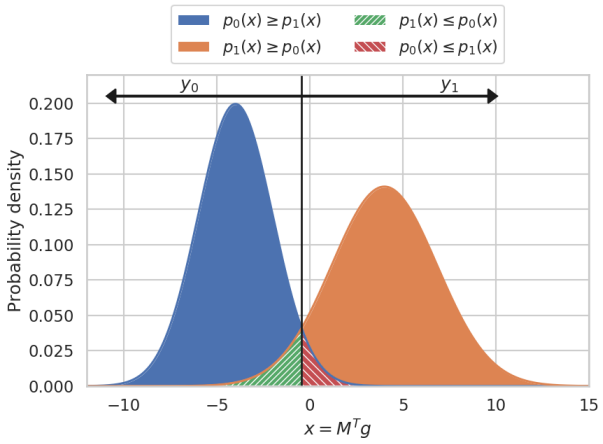

FIG. 1. *Section 2.1.1: Example of the pdf of two classes (0 and 1) in the projected space. Here, $\pi_0 = \pi_1 = \frac{1}{2}$ and $\mu(S_2) = 0$.*

From the inequality shown on $\mu(A_S)$ in Proposition 2.5, many relations can be established with other metrics [16]. In this paper we compare the success rate measure

to the Hellinger distance (see 2.1.2) and the symmetrized Kullback-Leibler divergence (see 2.1.3).

2.1.2. Relation to the Hellinger distance. The Hellinger distance (see Definition 2.7) is a f-divergence [7]. For some studies the Hellinger distance is preferred to other common f-divergences as the Kullback-Leibler divergence or χ^2-divergence which are not metrics [38].

DEFINITION 2.7. *The Hellinger distance d_H between two probability distributions P and Q on S, with pdfs p and q respectively is:*

$$d_H^2(P,Q) \triangleq \frac{1}{2}\int_S (\sqrt{p(x)} - \sqrt{q(x)})^2 dx. \tag{2.19}$$

The following result can be established:

PROPOSITION 2.8. *Let P_0 and P_1 be two probability distributions on S and p_0 and p_1 the corresponding pdf. Then,*

$$\min(\pi_0,\pi_1) d_H^2(P_0,P_1) + \frac{1}{2} \leq \mu(A_S) \leq \max(\pi_0,\pi_1)\sqrt{2} d_H(P_0,P_1) + \frac{1}{2}, \tag{2.20}$$

where d_H is the Hellinger distance (see Definition 2.7) and $\mu(A_S)$ the success events measure defined previously in Equation 2.8.

The proof of Proposition 2.8 is immediate using the result of the Proposition 2.5 and the inequalities between the Hellinger distance and the total variation, proposed in [10].

2.1.3. Relation to the symmetrized Kullback-Leibler divergence. The Kullback-Leibler divergence (or relative entropy) [25] (see Definition 2.9) is a measure of the dissimilarity of a probability distribution to another. It reads:

DEFINITION 2.9. *The Kullback-Leibler divergence between two probability distributions P and Q on Ω, with pdf p and q is:*

$$D_{KL}(P||Q) \triangleq \int_\Omega p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx. \tag{2.21}$$

In many classification problems, for symmetry reasons, the symmetrized Kullback-Leibler divergence is introduced: $D_{SKL}(P,Q) \triangleq \frac{1}{2}(D_{KL}(P||Q) + D_{KL}(Q||P))$. Aiming at improving the classification, the maximization of the symmetrized Kullback-Leibler divergence is proposed [3, 28, 30, 37]. Hereafter, a result is proved relating the classification score Eq.(2.8) to the symmetrized Kullback-Leibler divergence.

PROPOSITION 2.10. *Let P_0 and P_1 be two probability distributions on S (see Definition of the set S in 2.2) with pdf p_0 and p_1.*

If $\log\left(\frac{p_0}{p_1}\right) \in L^\infty(S)$ and, moreover, $D_{KL}(p_i||p_j) < +\inf$ for $i \neq j, i, j = 0$ or 1 then, $\exists c > 0$ such that the following inequalities hold:

$$c\frac{\mu(A_S) - \frac{1}{2}}{\min(\pi_0,\pi_1)} \geq D_{SKL}(P_0,P_1) \geq 2\left(\frac{\mu(A_S) - \frac{1}{2}}{\max(\pi_0,\pi_1)}\right)^2, \tag{2.22}$$

Proof. The right hand side inequality is proved by making use of the Pinsker inequality [13]:

$$\delta_{TV}^2(P_0, P_1) \leq \frac{1}{2} D_{SKL}(P_0, P_1). \tag{2.23}$$

Then, using the inequality between the total variation distance and the measure of the success events, Eq. 2.5, we directly get:

$$D_{SKL}(P_0, P_1) \geq 2 \left(\frac{\mu(A_S) - \frac{1}{2}}{\max(\pi_0, \pi_1)} \right)^2. \tag{2.24}$$

To prove the inequality on the left hand side we consider the definition of the symmetrized Kullback-Leibler divergence:

$$D_{SKL}(P_0, P_1) = \frac{1}{2} \int_S (p_0 - p_1) \ln \left(\frac{p_0}{p_1} \right) dx. \tag{2.25}$$

As, $\log \left(\frac{p_0}{p_1} \right) \in L^\infty(S)$, Hölder inequality leads to:

$$D_{SKL}(P_0, P_1) \leq \frac{\|\log(p_0/p_1)\|_{L^\infty}}{2} \int_S |p_0 - p_1| dx. \tag{2.26}$$

In what follows we set: $c = \|\log(p_0/p_1)\|_{L^\infty}$. The definition of the total variation is inserted:

$$D_{SKL}(P_0, P_1) \leq c \delta_{TV}(P_0, P_1). \tag{2.27}$$

Then,

$$D_{SKL}(P_0, P_1) \leq c \frac{\mu(A_S) - \frac{1}{2}}{\min(\pi_0, \pi_1)}, \tag{2.28}$$

which concludes the proof. ∎

Under the hypothesis of Propostion 2.10, we clearly see that the minimum of $D_{SKL}(P_0, P_1)$ is 0 and it is reached for $\mu(A_S) = \frac{1}{2}$ (the minimum of $\mu(A_S)$). Moreorer, increasing the success rate is equivalent to inscreasing the value of the symmetrized Kullback-Leibler divergence.

2.2. Optimisation of the classification success rate. The method proposed consists in choosing an element of the Stiefel manifold to define the input of the classifier: $x = M^T g$. The goal is to optimise the score $\mu(A_S)$ introduced and commented in the section above. Optimising over all the possible elements of the Stiefel manifolds (of multiple and unknown dimension k) would be prohibitive. To circumvent this, a double greedy approach is proposed. The heuristics are the following: the smaller the dimension of the input, the better it is in terms of palliating the curse of dimensionality; aiming at reducing possible overfitting phenomena, the sparser the orthonormal vectors of M, the better it is. Henceforth, the strategy which is investigated is the following: we start with $k = 1$ and look for a vector of unitary norm such that at each step of a greedy method, we maximise $\mu(A_S)$. When the error on a validation set stagnates and start increasing (early stopping criterion [36]), we start considering $k = 2$. The first column vector of M is the result of the previous step of the method, and by a greedy approach we construct a second unitary norm column vector, orthogonal to the first one. This can be iterated until the error on a validation set starts increasing as soon as we start building the $(k+1)-$th vector.

2.2.1. Computation of $\mu(A_S)$. Before detailing the double greedy algorithm in Section 2.2.2, let us introduce a strategy to approximate the measure of the success events $\mu(A_S)$. In general, the densities p_0, p_1 are not known. Instead, samples are given. To approximate the integral in Eq.(2.8), we use a Montecarlo approach: in the present case, it turns out to be a counting of how many samples are correctly classified (*i.e.* $y = y_*$). The difficulty is to precisely estimate the regions S_0, S_1, S_2. For that, an estimation of the values of p_0, p_1 is required. Since the dimension k is usually small (for instance $k = 1, 2, 3$), a Kernel Density Estimation (KDE) is a viable way to estimate the values of p_0, p_1 and hence to have an approximation of the decomposition of S. For larger values of k, KDE could become unpractical and costly from a numerical point of view [26]. A surrogate is proposed, based on the use of the Mahalanobis distance [11, 43]. This provides a perfect outcome in the case of Gaussian distributions. Since, in general, the projected densities p_0, p_1 are not Gaussians, an approximation based on hierarchical clustering is proposed. Roughly speaking, classes i ($i = 0, 1$) may be seen as a mixture of Gaussian distributions of means $(\mu_i^{(1)}, \dots, \mu_i^{(l)})$ and covariance matrices $(\Sigma_i^{(1)}, \dots, \Sigma_i^{(l)})$, that can be computed by clustering. For an observation x with a label $y_* = i$ the success event s is given by:

$$d_i^{(k)} = (x - \mu_i^{(k)})^T [\Sigma_i^{(k)}]^{-1} (x - \mu_i^{(k)}), \ i = 0, 1, \quad (2.29)$$

$$s(x) = \begin{cases} 1 \text{ if } \min_{k=1,\dots,l_i} d_i^{(k)} < \min_{k=1,\dots,l_j} d_j^{(k)} \text{ and } y_* = i \ (j \neq i) \\ 0 \text{ otherwise} \end{cases} . \quad (2.30)$$

For all the entries of the dataset, the individual score s proposed in Eq.(2.29) can be evaluated. The approximation of $\mu(A_S)$ to be used reads:

$$\mu(A_S) \approx \frac{\pi_0}{n_0} \sum_{l=1}^{n_0} s(x^{(l)}_{|y_*=0}) + \frac{\pi_1}{n_1} \sum_{l=1}^{n_1} s(x^{(l)}_{|y_*=1}), \quad (2.31)$$

where $n_0 + n_1 = n_s$, with n_0 and n_1 are the number of samples labeled $y_* = 0$ and $y_* = 1$ respectively. This is an empirical approximation of the score introduced in Eq.(2.8). The error introduced by such an approximation and possible alternatives are discussed in [4].

An example of score estimation is shown in Figure 2. In this example, the distribution of the class 0 is a mixture of a Gaussian and a uniform distribution; class 1 is a mixture of two Gaussian distributions. Samples are drawn and the hierarchical clustering algorithm applied.

The bound of the probability of being in class i is then given by the multivariate Chebyshev inequality [33].

2.2.2. Double greedy algorithm. Let $n_s, n_v \in \mathbb{N}^*$ be the number of the samples used in the training and the validation phases respectively. A training and a validation datasets $(g^{(i)}, y_*^{(i)})_{i=1}^{n_s}$, $(g^{(i)}, y_*^{(i)})_{i=1}^{n_v}$ are given, that consist of couples of dictionary entries and corresponding labels.

Let $\widehat{M}_{k,n_g} \in \mathcal{M}_{k,n_g}$ be the element of the Stiefel Manifold selected at the $k-$th outer iteration of the method. The goal is to find a vector $\omega_* \in \mathbb{R}^{n_g}$, orthogonal to

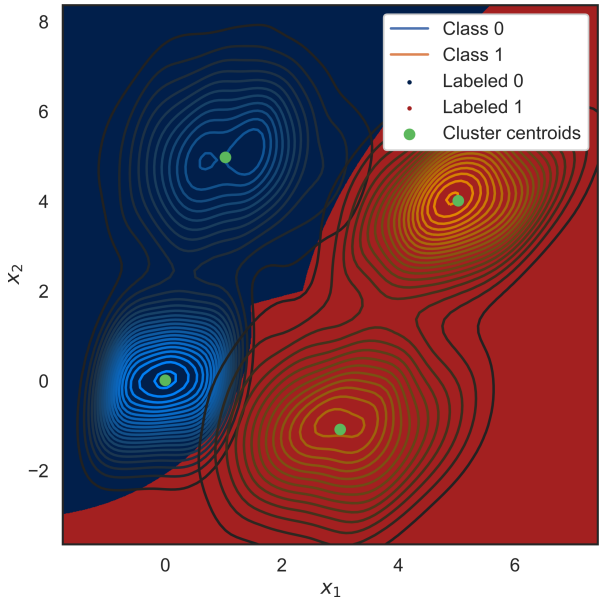

FIG. 2. *Section 2.2.1: Example of classification using Mahalanobis distance. The Kernel Density Estimation using a Gaussian kernel shows the distribution for the two classes. Cluster centroids were obtained using DBSCAN. Class 0: uniform distribution on the square centered on 0 and a side of length 1 and a Gaussian bivariate distribution with* $\mu = (1, 5)$ *and identity covariance matrix. Class 1: Gaussian bivariate distributions* $\mu_a = (5, 4)$ *and* $\mu_b = (3, -1)$ *with* $\Sigma_a = (0.8, 0.2; 0.2, 0.6)$ *and* $\Sigma_b = Id$*. Sample size of 500 for each distribution.*

all the columns of the matrix $\hat{M}_{k,n_g}$, such that:

$$\widehat{M}_{k+1,n_g} = [\widehat{M}_{k,n_g}, \omega_*], \tag{2.32}$$

$$x \in \mathbb{R}^{k+1}, \ x = \widehat{M}^T_{k+1,n_g} g, \tag{2.33}$$

$$\omega_* = \arg \inf_{\omega \in \mathbb{R}^{n_g}} \mu(A_S). \tag{2.34}$$

When n_g is large, this optimisation can be costly. Furthermore, when the vector ω is sparse the classification tends to be less prone to overfitting phenomena. For these reasons, ω is constructed in a greedy way. At first $\|\omega\|_{\ell^{0,n_g}} = 1$, so that only one dictionary entry is chosen, by computing the value of the score (on the training dataset) for all possible choices and keeping the best.

At the beginning of the $l-$th inner iteration, $\|\omega\|_{\ell^{0,n_g}} = l - 1$, $l - 1$ dictionary entries have been chosen and we have to choose the $l-$th one. Let the chosen indices be in the set $c^{(k+1)} = \{i_1, \ldots, i_{l-1}\}$. The $l-$th non zero entry has to be chosen among the indices $i \in c_c^{(k+1)}$, the complementary set of $c^{(k+1)}$. Moreover, the best values of the selected entries of ω are sought, such that the result of the classification is the best possible (in the sense of the score introduced). Once one candidate to be the $l-$th non zero component is proposed, an optimisation task on the entries of ω is performed by using the CMAES method, detailed in [22, 23]. This does not guarantee automatically that ω is orthogonal to the subspace spanned by the column of $\widehat{M}_{k,n_g}$. Otherwise stated, $[\widehat{M}_{k,n_g}, \omega] \in Gr_{k+1,n_g}$. The projection onto the Stiefel manifold is

obtained by QR decomposition. Let $Q_m \in \mathbb{R}^{n_g \times k+1}$, $R_m \in \mathbb{R}^{k+1 \times k+1}$, it holds:

$$Q_m R_m = [\widehat{M}_{k,n_g}, \omega], \tag{2.35}$$

$$\widehat{M}_{k+1,n_g} = Q_m. \tag{2.36}$$

Among all the possible optimised choices for the $l-$th component, the one that maximises the score is chosen. As said, the stopping criterion for these iterations is the early stopping strategy [36]: the score is computed on the validation set. A stagnation of the score ends the inner iteration. As soon as increasing the dimension of the Stiefel manifold does not produce an improvement on the score computed on the validation, the outer iterations end. Once the algorithm terminates, the element of the Stiefel manifold is obtained.

The details of the method are shown in Algorithm 2.2.2.

Remark. When, in the proposed method, the $\|\omega\|_{\ell^{0,n_g}} = 1$, $\forall k$, the Feature Selection (FS) [18] reduction is retrieved, as a particular case. Furthermore, when the objective function is not the quantity $\mu(A_S)$ but the ℓ^2 norm of the samples g reconstruction, the proposed approach turns out to be a sparse approximation of the Principal Component Analysis (PCA) of the data (a description is provided in [5, 42]). The outcome of the proposed method is therefore a set of orthonormal modes that does not coincide with the PCA modes. These two methods, FS and PCA, are the most used dimension reduction techniques when dealing with classification problems. A numerical comparison will be proposed in Section 3.

2.3. Principle of analysis. In this section an analysis of the proposed method is presented. The goal is to show that, in the limit case of an infinite number of samples, or, in alternative, the perfect knowledge of the pdf, the proposed method tends to maximise the score. In the case in which all the dictionary entries are used, the score by exploiting all the entries is retrieved.

PROPOSITION 2.11. *Let $\widehat{M}_{k,n_g} \in \mathcal{M}_{k,n_g}$ and $\widehat{M}_{k,n_g} = [\widehat{M}_{k-1,n_g}, \omega]$; let $1 \leq m < n_g$, and $\|\omega\|_{\ell^{0,n_g}} = m$. The set of non-zero entries of ω is denoted by c, whose cardinality is $\#c = m$. Let $\tilde{\omega} \in \mathbb{R}^{n_g}$. The set of non-zero entries of $\tilde{\omega}$ is $\tilde{c}$, $\#c = m+1$. It holds $c \subset \tilde{c}$. Then,*

$$\max_{\tilde{\omega}} \mu(A_S) \geq \max_{\omega} \mu(A_S). \tag{2.37}$$

Proof. The proof is immediate. The function $\mu(A_S)$ is continuous and bounded, the Stiefel manifold is a compact set, and the set of non-zero entries of ω is strictly included in the one of $\tilde{\omega}$. Henceforth, the conclusion. Indeed, at worst, the non-zero entry of $\tilde{\omega}$ which is not a non-zero entry of ω can be set to zero and the equality would hold. ∎

This proposition shows that, in the inner iteration, as far as we add terms to the vector ω, the score improves.

The outer iteration, the one in which the dimension of the element of the Stiefel manifold is increased, is the object of the following Proposition.

PROPOSITION 2.12. *Let $M_{k,n_g} \in \mathcal{M}_{k,n_g}$ and the associated score be $\mu(A_S^{(k)})$. Let $M_{k+1,n_g} \in \mathcal{M}_{k+1,n_g}$ such that:*

$$M_{k+1,n_g} = [M_{k,n_g}, \omega],$$

Algorithm 2.1 Double greedy algorithm

$k \leftarrow 1$ ▷ Dimensional counter.
$\mu(A_S)_v^{new} \leftarrow 1/2$ ▷ Minimal reachable value of $\mu(A_S)$ for the validation set.
$\mu(A_S)_v^{old} \leftarrow 0$ ▷ Success event measure of the validation set. See†.
$c \leftarrow [1, \dots, n_g]$ ▷ Dictionary entry indices.
$\widehat{M} \leftarrow [\,]$ ▷ Empty matrix which will be an element of $\mathbf{Gr}_{k,n_g}$.
while $\mu(A_S)_v^{new} > \mu(A_S)_v^{old}$ **do** ▷ Loop on the dimension.
  $j \leftarrow 1$ ▷ Non-zero component counter.
  $c^{(k)} \leftarrow [\,]$ ▷ Empty vector which stores dictionary entry indices.
  while $\mu(A_S)_v^{new} > \mu(A_S)_v^{old}$ **do** ▷ Loop on the non-zero components.
    $\mu(A_S)_v^{old} \leftarrow \mu(A_S)_v^{new}$ ▷ Update stop criterion.
    $\mu \leftarrow [0]^{n_g}$ ▷ Zero vector of size n_g which stores success event measures.
    $W \leftarrow [0]^{n_g \times n_g}$ ▷ Empty matrix of size $n_g \times n_g$ which stores weights.
    $c_c^{(k)} \leftarrow c \backslash c^{(k)}$
    for $l \in c_c^{(k)}$ **do**
      Initialize ω_l ▷ Initialize non-zeros indices for CMAES.
      $\mu(A_S)_l, \omega_l \leftarrow CMAES(\omega_l, (g_i, y_i^*)_{i=1}^{n_s})$ ▷ Optimisation over ω_l. See††.
      $\mu_l \leftarrow \mu(A_S)_l$ ▷ Assign the l^{th} component of μ.
      $\omega \leftarrow Weights(\omega_l, s_j, l)$ ▷ Generate l^{th} weight column vector of W. See‡.
      $W_l \leftarrow \omega$ ▷ Assign the l^{th} column of the weight matrix W.
    end for
    $l_* \leftarrow \mathrm{argmax}_l\, \mu_l$ ▷ New dictionary entry position for the contribution.
    $\omega_* \leftarrow W_{l_*}$ ▷ Extract corresponding weights.
    $\widehat{M}_* \leftarrow [\widehat{M}, \omega_*]$
    $M_* \leftarrow QR(\widehat{M}_*)$
    $\mathcal{D}_v \leftarrow (M_*^T g_i, y_i^*)_{i=1}^{n_v}$ ▷ Compute the projected validation dataset.
    $\mathcal{D}_t \leftarrow (M_*^T g_i, y_i^*)_{i=1}^{n_v}$ ▷ Compute the projected training dataset.
    $\mu(A_S)_v^{new} \leftarrow Score(\mathcal{D}_v, \mathcal{D}_t)$ ▷ Compute score on the validation set‡‡:.
    $s_j \leftarrow [s_j, l_*]$
    $j \leftarrow j + 1$
  end while
  $\widehat{M} \leftarrow [\widehat{M}, \omega_*]$
  $k \leftarrow k + 1$
end while
  return $\widehat{M}$

†: Any value lower than $\mu(A_S)_v^{new}$ to enter in the while loop.
††: For each ω_l computed at each CMAES steps, the QR decomposition of $[\widehat{M}, \omega_l]$ and projection of the taining set $(x_i = M^T g_i, y_i^*)_{i=1}^{n_s}$ are performed to compute $\mu(A_S)_l$.
‡:$[0]^{n_g}$ vector with optimised weights assigned to the non-zero positions s_j and l.
‡‡: The validation score is computed using KDE or Mahalanobis distance through the projected training set.

where $\omega \in \mathbb{R}^{n_g}$ *and the associated score be* $\mu(A_S^{(k+1)})$. *Then:*

$$\mu(A_S^{(k+1)}) \geq \mu(A_S^{(k)}). \tag{2.38}$$

Proof. Let us denote $h = \pi_0\rho_0 - \pi_1\rho_1$, S_k the space obtained by projecting $g \in \mathbb{R}^{n_g}$ onto the columns of M_{k,n_g}. Let its orthogonal complement be denoted by $S_k^{\perp}$. The element of the orthogonal group constructed from M_{k,n_g} is denoted by $R = [M_{k,n_g}, M_{k,n_g}^{\perp}]$. It holds:

$$\xi = R^T g, \tag{2.39}$$

$$x = [\xi_1; \dots; \xi_k]. \tag{2.40}$$

As remarked in Eq.(2.2) $p(x)$ is obtained by:

$$p(x) = \int_{S_k^{\perp}} \rho(\xi) d\xi_1, \dots d\xi_k. \tag{2.41}$$

The score (and the total variation) is then directly related to the following integral:

$$I^{(k)} = \int_{S_k} \left| \int_{S_k^{\perp}} h \; d\xi_{k+1} \dots d\xi_{n_g} \right| d\xi_1 \dots d\xi_k. \tag{2.42}$$

Without loss of generality let us suppose that:

$$\xi_{k+1} = \omega^T g. \tag{2.43}$$

Remark that the orthogonal complement to S_k can be always constructed in this way. We will denote by $S_{k+1}^{\perp} = S_k^{\perp}/\xi_{k+1}$. Hence:

$$I^{(k)} = \int_{S_k} \left| \int_{-\infty}^{\infty} \left(\int_{S_{k+1}^{\perp}} h \; d\xi_{k+2} \dots d\xi_{n_g} \right) d\xi_{k+1} \right| d\xi_1 \dots d\xi_k. \tag{2.44}$$

When ω is used to construct the input ($x = M_{k+1,n_g}$), the integral $I^{(k+1)}$ reads:

$$I^{(k+1)} = \int_{S_k} \int_{-\infty}^{\infty} \left| \int_{S_{k+1}^{\perp}} h \; d\xi_{k+2} \dots d\xi_{n_g} \right| d\xi_{k+1} d\xi_1 \dots d\xi_k. \tag{2.45}$$

A straightforward inequality follows:

$$\int_{-\infty}^{\infty} \left| \int_{S_{k+1}^{\perp}} h \; d\xi_{k+2} \dots d\xi_{n_g} \right| \; d\xi_{k+1} - \left| \int_{-\infty}^{\infty} \left(\int_{S_{k+1}^{\perp}} h \; d\xi_{k+2} \dots d\xi_{n_g} \right) d\xi_{k+1} \right| \geq 0, \tag{2.46}$$

which implies:

$$\mu(A_S^{(k+1)}) - \mu(A_S^{(k)}) = \frac{1}{2} \left(I^{(k+1)} - I^{(k)} \right) \geq 0, \tag{2.47}$$

and this concludes the proof. ∎

Remark. Since, at each step of the method, we enforce that the matrices M_{k,n_g} belong to the Stiefel manifold, when $k = n_g$ we retrieve an element of the orthogonal group, whose associated score is the maximal possible.

3. Computational studies. In this section, we compare the algorithm with classical tools used for dimension reduction in the context of classification problems. The first part consists of comparing the strategy proposed in this paper with Feature Selection (FS) [18]. In the second part we make the comparison with the Principal Component Analysis (PCA) method [5, 42].

3.1. Comparison with feature selection. FS is a widely used dimension reduction tool consisting of selecting a subset of features, pertinent to answer a clustering [9] or classification [8] problem. In the context of the present work, this would consist in selecting a subset of the dictionary entries, and, as remarked, it can be seen as a particular case of the proposed method.

In this first test case a synthetic example is constructed by considering a gaussian mixture: $\rho_0(g), \rho_1(g)$ are two normal distributions, of mean and variance (μ_0, Σ_0) and (μ_1, Σ_1) respectively. When dealing with Gaussian distributions, the symmetrized Kullback-Leibler divergence can be analytically computed. In Section 2.1, an equivalence between the symmetrized KL divergence and the score $\mu(A_S)$ is shown. The symmetrized Kullback-Leibler divergence between the distributions reads:

(3.1)
$$D_{SKL}(\rho_0, \rho_1) = \frac{1}{4}\left(tr(\Sigma_1^{-1}\Sigma_0 + \Sigma_0^{-1}\Sigma_1) + (\mu_1 - \mu_0)^T(\Sigma_0^{-1} + \Sigma_1^{-1})(\mu_1 - \mu_0) - 2n_g\right).$$

Let $k \in \mathbb{N}^*$ denote the number of entries selected by the FS, let $l \in \mathbb{N}^*, l \leq n_g$ be such that the elements of the Stiefel manifold are $(M_{l,n_g} \in \mathcal{M}_{l,n_g})$ and $m \in \mathbb{N}^*$ be the maximal number of non-zero entries of the columns of M_{l,n_g}.

When projecting Gaussian distributions on linear subspaces, Gaussian distributions are retrieved, namely P_0, P_1, whose densities are $p_0(x), p_1(x)$. The mean and variances of these are reported in Table 1 for the case of FS and the proposed method.

Classification strategy	Σ_0	Σ_1	μ_0	μ_1
Feature Selection	I_k	$\beta I_k, \beta > 0$	0_k	1_k
Double Greedy Algorithm	I_l	$\beta I_l, \beta > 0$	0_l	$(\sqrt{m}, \ldots, \sqrt{m}) \in \mathbb{R}^l_+$

TABLE 1

Section 3.1: Gaussian parameters for feature selection and double greedy algorithm study case.

The symmetrized Kullback-Leibler divergence for FS and, respectively, for the double greedy algorithm (DGA) reads:

$$D_{SKL}^{(FS)}(P_0, P_1) = \frac{1}{4}\left(\frac{k}{\beta} + k\beta + k(1 + \frac{1}{\beta}) - 2k\right), \tag{3.2}$$

$$D_{SKL}^{(DGA)}(P_0, P_1) = \frac{1}{4}\left(\frac{l}{\beta} + l\beta + lm(1 + \frac{1}{\beta}) - 2l\right). \tag{3.3}$$

An analysis of the above expressions provides some insight on the performances of the methods. Let us consider the difference between the divergences:

$$f_{l,k}(\beta) = \left(\frac{k}{l} - 1\right)\beta^2 + \left(2 - \frac{k}{l}\right)\beta + 2\frac{k}{l} - 1, \tag{3.4}$$

$$\Delta = D_{SKL}^{(DGA)}(P_0, P_1) - D_{SKL}^{(FS)}(P_0, P_1) \geq 0 \iff m \geq \frac{f_{l,k}(\beta)}{\beta + 1}. \tag{3.5}$$

Some properties are highlighted:

- If $k = l$, then, $\forall\beta$, the symmetrized KL divergence is larger for DGA if $m \geq 1$; in the case in which $m = 1$, as commented before, the methods coincide.
- If $k < l$, $\forall\beta$, $\Delta \geq 0$: in this case DGA always outperforms FS.
- if $k > l$, different scenarios are possible.
- It is interesting to consider the case of identical Gaussians, namely $\beta = 1$, the DGA outperforms FS if $m > \frac{k}{l}$. Remark that when $l = 1$ (DGA selects just an element of the unit sphere): $\Delta \geq 0$ if $m \geq k$.

In general, when both the methods achieve the same result in terms of symmetrized KL divergence, DGA method has a reduced dimension smaller (in some cases much smaller) than FS. This is particularly relevant when a finite (and not so large) number of samples are available. A comparison is given in Table 2 where the symmetrized Kullback-Leibler difference between DGA and FS is computed for some values of k, l and m.

Δ		$m=1$					$m=5$					$m=10$				
		l					l					l				
		1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
k	1	0.0	2.0	4.5	7.0	9.5	2.0	12.0	24.5	37.0	49.5	4.5	24.5	49.5	74.5	99.5
	5	-2.0	0.0	2.5	5.0	7.5	0.0	10.0	22.5	35.0	47.5	2.5	22.5	47.5	72.5	97.5
	10	-4.5	-2.5	0.0	2.5	5.0	-2.5	7.5	20.0	32.5	45.0	0.0	20.0	45.0	70.0	95.0
	15	-7.0	-5.0	-2.5	0.0	2.5	-5.0	5.0	17.5	30.0	42.5	-2.5	17.5	42.5	67.5	92.5
	20	-9.5	-7.5	-5.0	-2.5	0.0	-7.5	2.5	15.0	27.5	40.0	-5.0	15.0	40.0	65.0	90.0

TABLE 2

Secion 3.1: Δ difference between symmetrized Kullback-Leibler divergences obtained using DGA (parameters l and m) and FS (parameter k) defined in Equation 3.4. Here, the covariance matrix factor β is set to 1 (see Table 1).

3.2. Comparison with PCA. In this section, we compare the proposed double greedy algorithm with the Principal Component Analysis (PCA), which consists in finding an orthogonal transformation such that the variance of the dataset in the principal directions is the largest possible [5, 42].

Let G follows a multivariate normal distribution in $\mathbb{R}^{n_g}$. Let $l \in \mathbb{N}^*$ and $I = \{i_1, \ldots, i_l\}$ be a set of indices. For this test case, classes $0, 1$ are defined as:

$$\begin{cases} y^* = 0 \text{ if } g_{i_1}, \ldots, g_{i_l} \geq 0, \\ y^* = 1 \text{ otherwise.} \end{cases} \tag{3.6}$$

In the particular case analysed hereafter, $n_g = 50$ and $I = \{10, 11, 12, 13\}$. The total number of samples in the training set is $n_s = 1500$, 105 for the class 0 and 1395 for the class 1. The number of samples for the validation set is $n_v = 500$, 35 samples for the class 0 and 465 for the class 1. The method proposed used $m = 2$ dictionary entries to build the first direction and $m = 3$ dictionary entries to build the second direction.

In Figure 3, the samples projected on the first two principal directions obtained by PCA are shown. As it can be assessed, PCA does not provide an efficient preprocessing, the conditional densities of the classes being practically indistinguishable.

On the other hand, the two first directions identified by the Double Greedy method proposed (M_{2,n_g}) tend to maximise the separation between the conditional densities. The samples projected on these directions are shown in Figure 4.

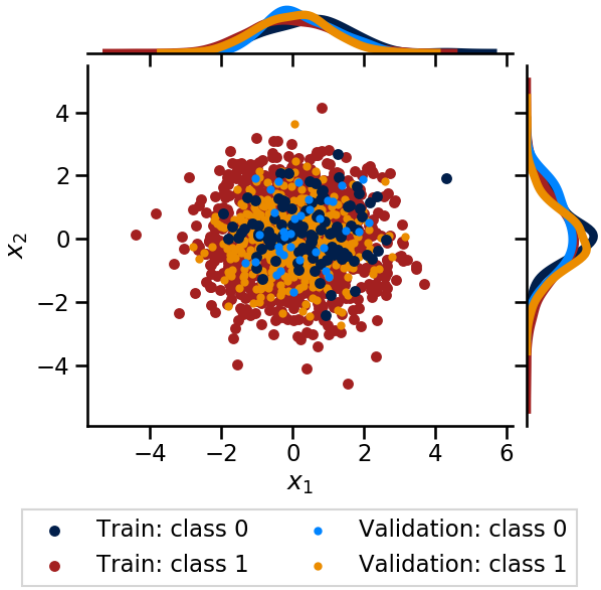

FIG. 3. *Section 3.2: samples projected on the two first directions computed by PCA, along with the marginal conditional densities.*

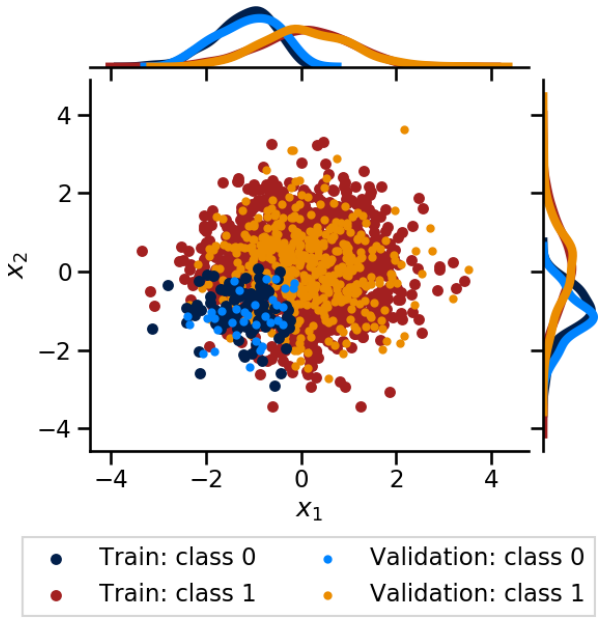

FIG. 4. *Section 3.2: samples projected on M_{2,n_g} obtained by the double greedy approach and the associated marginal conditional densities. The Mahalanobis distance was used for the classification (see Section 2.2.1). Using the early-stopping criterion on the validation set, two components were chosen for the first dimension and three for the second.*

The results obtained allow to stress an important aspect. PCA is a general purpose reduction method, which is often effective, but it is not specific to classification tasks, as the method proposed. Henceforth, there are situations, like the one shown in

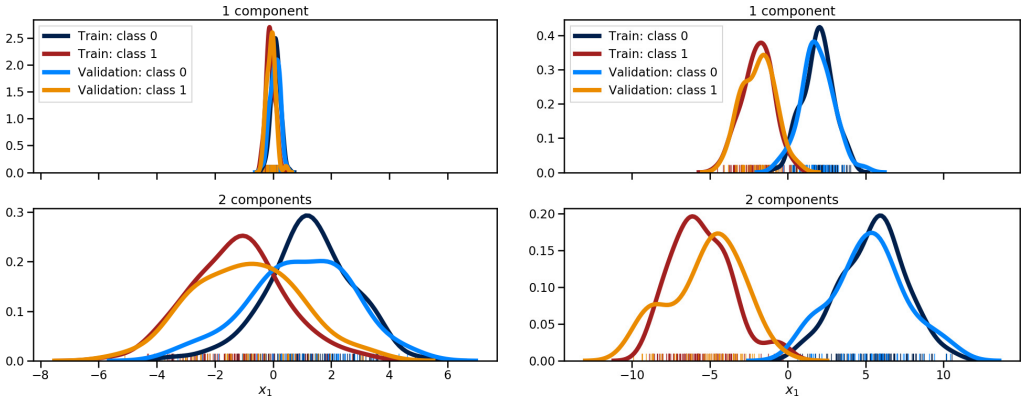

FIG. 5. *Test case in Section 3.3, projected distributions on $x \in \mathbb{R}$ for the case: (a) $\eta = 1$ and (b) $\eta = 4$. Upper row, $\|\omega\|_{\ell^{0,n_g}} = 1$, lower, $\|\omega\|_{\ell^{0,n_g}} = 2$.*

this example, in which PCA fails in providing a well performing dimension reduction.

3.3. A high-dimensional low sample size example. We consider a numerical illustration of a high-dimensional low sample size regime. For this $n_g = 10^5$, and the number of samples in the training set is $n_s = 200$, evenly distributed between the two classes. The validation set consists of $n_v = 100$ samples. Let $I = 10, \dots, 20$ be a set of indices, whose cardinality is $\#I = 11$. The probability density function reads:

$$\rho(g) = \frac{1}{2}\rho_0(g) + \frac{1}{2}\rho_1(g), \tag{3.7}$$

where ρ_0, ρ_1 are unitary variance Gaussians, whose mean are $\mu_0 = [0, \dots, 0]$ and:

$$\begin{cases} \mu_{1i} = \eta \text{ if } i \in I, \\ \mu_{1i} = 0 \text{ otherwise.} \end{cases} \tag{3.8}$$

For this example, we considered two cases, namely $\eta = 1$ and $\eta = 4$; the Mahalanobis distance criterion was used to approximate the score.

The results are shown in Fig. 5, when the dimension of the reduced space is $k = 1$. The samples are projected in $x \in \mathbb{R}$ and the probability densities of the two classes are plotted for the training and validation sets. In the upper row, $\|\omega\|_{\ell^{0,n_g}} = 1$, in the lower row $\|\omega\|_{\ell^{0,n_g}} = 2$. Visually, we can assess that the separation between the densities increases when we use two components instead of one, and this holds for both the training and the validation sets; this is confirmed by the increase in the classification score. When $\eta = 4$, the densities of the two classes ρ_0, ρ_1 have a larger total variation with respect to the case $\eta = 1$. This is found also for the marginal densities p_0, p_1. The two non-zero components chosen by the algorithm to construct the first direction (M_{1,n_g}) are elements of I described above. Selected non-zero components are 12 and 10 for $\eta = 1$ study case and 14 and 11 for $\eta = 4$.

3.4. Application to a classification problem. To conclude, we present a test case based on a realistic dataset: the LSVT Voice Rehabilitation data set provided by UCI machine learning repository (https://archive.ics.uci.edu/ml/index.php). This

dataset based on dysphonia measures is studied to assess the LSVT protocol in Parkinson disease [41], and was used, as other datasets in this repository, as a benchmark to test several classification strategies.

We randomly divided the dataset into a training set of size $n_s = 50$ and a validation set of size $n_v = 76$. The feature space, which is, in our case, the set of the dictionary entries, is $n_g = 309$.

The following comparison is proposed: the PCA and the proposed (DGA) method are applied to define a low dimensional input, that will be used to perform the classification with several standard methods.

Concerning the PCA, we choose the $k = 3$ first directions (which explain approximately 99.8% of the variance). The double greedy algorithm (DGA), using an early stopping criterion on the validation set, stopped at $k = 2$, with one component ($m = 1$) for the first direction and two components ($m = 2$) for the second direction.

The chosen classifiers for the training and validations steps are Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Tree (DT), Naive Bayes (NB) and Support Vector Machines (SVM). We used the Scikit-learn library [35] with default parameters.

The success rates on the validation set for the two dimension reduction strategies and the five classifiers are given in Table 3.

	LDA	KNN	DT	NB	SVM
PCA	0.58	0.49	0.59	0.51	0.67
DGA	0.76	0.74	0.78	0.82	0.68

TABLE 3

Section 3.4: Validation dataset success rate using PCA and DGA strategies for dimension reduction. Classifiers used with their default parameters with the Scikit-learn [35] library.

For all the tested classifiers, the DGA strategy shows an increase in the validation dataset success rate. Except for SVM, this gain is particularly significant.

4. Conclusions. This paper investigates a double greedy algorithm to construct the input x of a classifier by exploiting a large number of dictionary entries. The method is designed to deal with classification problems in a high-dimensional/low sample size regime. The method can be interpreted as a sparse goal oriented dimension reduction technique. The first contribution is the introduction of an objective function to be maximised, which is directly related to the performances of the classifiers in the reduced space. This objective function was related to quantities which are commonly used to assess the performances in classification problems. The method proposed is easily parallelisable and hence well adapted to large problems. Some examples are proposed to illustrate the performances of the proposed method: first, a comparison in a small scale problem is performed with Feature Selection and the Principal Component Analysis; then, the method was tested on a large scale synthetic example that mimics a high-dimensional/low sample size regime and a realistic dataset.

Several perspectives arise. One concerns the application of the method to a broader set of realistic cases. The extension to more than two classes as well as to regression problems will be considered.

REFERENCES

[1] A. R. Barron, L. Gyorfi, and E. C. van der Meulen, *Distribution estimation consistent in total variation and in two types of information divergence*, IEEE transactions on Information Theory, 38 (1992), pp. 1437–1454.
[2] R. E. Bellman, *Adaptive control processes: a guided tour*, vol. 2045, Princeton university press, 2015.
[3] B. Bigi, *Using kullback-leibler distance for text categorization*, in European Conference on Information Retrieval, Springer, 2003, pp. 305–319.
[4] P. Binev, A. Cohen, W. Dahmen, R. DeVore, et al., *Classification algorithms using adaptive partitioning*, The Annals of Statistics, 42 (2014), pp. 2141–2163.
[5] C. M. Bishop, *Pattern recognition and machine learning*, springer, 2006.
[6] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, and D. Picard, *Capturing ridge functions in high dimensions from point queries*, Constructive Approximation, 35 (2012), pp. 225–243.
[7] I. Csiszár, *Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten*, Magyer Tud. Akad. Mat. Kutato Int. Koezl., 8 (1964), pp. 85–108.
[8] M. Dash and H. Liu, *Feature selection for classification*, Intelligent data analysis, 1 (1997), pp. 131–156.
[9] M. Dash and H. Liu, *Feature selection for clustering*, in Pacific-Asia Conference on knowledge discovery and data mining, Springer, 2000, pp. 110–121.
[10] C. Daskalakis and Q. Pan, *Square hellinger subadditivity for bayesian networks and its applications to identity testing*, arXiv preprint arXiv:1612.03164, (2016).
[11] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, *The mahalanobis distance*, Chemometrics and intelligent laboratory systems, 50 (2000), pp. 1–18.
[12] L. Dümbgen, P. Del Conte-Zerial, et al., *On low-dimensional projections of high-dimensional distributions*, in From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner, Institute of Mathematical Statistics, 2013, pp. 91–104.
[13] A. A. Fedotov, P. Harremoës, and F. Topsoe, *Refinements of pinsker's inequality*, IEEE Transactions on Information Theory, 49 (2003), pp. 1491–1498.
[14] I. K. Fodor, *A survey of dimension reduction techniques*, tech. report, Lawrence Livermore National Lab., CA (US), 2002.
[15] M. Fornasier, K. Schnass, and J. Vybiral, *Learning functions of few arbitrary linear parameters in high dimensions*, Foundations of Computational Mathematics, 12 (2012), pp. 229–262.
[16] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, International statistical review, 70 (2002), pp. 419–435.
[17] N. Gunduz and E. Fokoué, *Robust classification of high dimension low sample size data*, arXiv preprint arXiv:1501.00592, (2015).
[18] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, Journal of machine learning research, 3 (2003), pp. 1157–1182.
[19] P. Hall, J. S. Marron, and A. Neeman, *Geometric representation of high dimension, low sample size data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 427–444.
[20] P. Howland, M. Jeon, and H. Park, *Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition*, SIAM Journal on Matrix Analysis and Applications, 25 (2003), pp. 165–179.
[21] G. Hughes, *On the mean accuracy of statistical pattern recognizers*, IEEE transactions on information theory, 14 (1968), pp. 55–63.
[22] M. W. Iruthayarajan and S. Baskar, *Evolutionary algorithms based design of multivariable pid controller*, Expert Systems with applications, 36 (2009), pp. 9159–9167.
[23] S. Kern, S. D. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos, *Learning probability distributions in continuous evolutionary algorithms–a comparative review*, Natural Computing, 3 (2004), pp. 77–112.
[24] H. Kim, H. Park, and H. Zha, *Distance preserving dimension reduction for manifold learning*, in Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, 2007, pp. 527–532.
[25] S. Kullback and R. A. Leibler, *On information and sufficiency*, The annals of mathematical statistics, 22 (1951), pp. 79–86.
[26] N. Langrené and X. Warin, *Fast and stable multivariate kernel density estimation by fast sum updating*, Journal of Computational and Graphical Statistics, (2019), pp. 1–27.
[27] B. Liu, Y. Wei, Y. Zhang, and Q. Yang, *Deep neural networks for high dimension, low*

sample size data., in IJCAI, 2017, pp. 2287–2293.
[28] C. Liu and H.-Y. Shum, *Kullback-leibler boosting*, in 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., vol. 1, IEEE, 2003, pp. I–I.
[29] Y. Liu, D. N. Hayes, A. Nobel, and J. S. Marron, *Statistical significance of clustering for high-dimension, low–sample size data*, Journal of the American Statistical Association, 103 (2008), pp. 1281–1293.
[30] M. Lopes, M. Fauvel, S. Girard, and D. Sheeren, *High dimensional kullback-leibler divergence for grassland management practices classification from high resolution satellite image time series*, in 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2016, pp. 3342–3345.
[31] S. Mayer, T. Ullrich, and J. Vybiral, *Entropy and sampling numbers of classes of ridge functions, vol. 42*, 2015.
[32] E. Meckes, *Approximation of projections of random vectors*, Journal of Theoretical Probability, 25 (2012), pp. 333–352.
[33] J. Navarro, *A simple proof for the multivariate chebyshev inequality*, arXiv preprint arXiv:1305.5646, (2013).
[34] I. Nourdin and G. Poly, *Convergence in law implies convergence in total variation for polynomials in independent gaussian, gamma or beta random variables*, in High Dimensional Probability VII, Springer, 2016, pp. 381–394.
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.
[36] L. Prechelt, *Automatic early stopping using cross validation: quantifying the criteria*, Neural Networks, 11 (1998), pp. 761–767.
[37] J. Ramírez, J. C. Segura, C. Benítez, A. De La Torre, and A. J. Rubio, *A new kullback-leibler vad for speech recognition in noise*, IEEE signal processing letters, 11 (2004), pp. 266–269.
[38] A. Shemyakin et al., *Hellinger distance and non-informative priors*, Bayesian Analysis, 9 (2014), pp. 923–938.
[39] S. Smale and D.-X. Zhou, *Estimating the approximation error in learning theory*, Analysis and Applications, 1 (2003), pp. 17–41.
[40] G. V. Trunk, *A problem of dimensionality: A simple example*, IEEE Transactions on Pattern Analysis & Machine Intelligence, (1979), pp. 306–307.
[41] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, *Objective automatic assessment of rehabilitative speech treatment in parkinson's disease*, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 22 (2013), pp. 181–190.
[42] S. Wold, K. Esbensen, and P. Geladi, *Principal component analysis*, Chemometrics and intelligent laboratory systems, 2 (1987), pp. 37–52.
[43] S. Xiang, F. Nie, and C. Zhang, *Learning a mahalanobis distance metric for data clustering and classification*, Pattern recognition, 41 (2008), pp. 3600–3612.